How can we push machine learning models – which have demonstrated impressive general-purpose capabilities in recent times – to be usable as reliable tools and agents? Furthermore, as a scientific community, how can we develop our systems to be efficient, low-cost, and accessible to researchers? I have been trying to answer these questions as a student, and now research assistant, at Princeton University working with Prof. **Kai Li** and Prof. **Karthik Narasimhan**, as well as part-time at Snapchat Research with Dr. **Neil Shah**. I am particularly passionate about research in **ML systems, natural-language processing, and applications of AI to the natural sciences**, and I have primarily been working on (1) building **efficient** systems and modules, (2) understanding and **evaluating model capabilities** in a rigorous and theoretically-motivated way, and (3) developing natural **agentic** frameworks for language models.

I recently graduated as **the top computer science (CS) student** in my Princeton undergraduate class (Phillip Goldman '86 Senior Prize), with a coursework centered around CS & probability theory, systems programming, and AI. Throughout the last four years, I have been fortunate enough to **publish and research on topics** in language agents / language + RL [1, 2, 3], model evals [4, 2], CS theory [5], and efficient deep learning[1] [2] through the Princeton Language and Intelligence group (PLI) and Prof. Kai Li's systems for ML lab. I also frequently **contribute to open-source ML research**, most recently creating novel fused Triton/CUDA kernels for Ligo Bioscience's open-source AlphaFold3[1] (900+ GitHub stars) to fit their model on a single device (technical report soon) and contributing to the GPU MODE "popcorn" project for training LLMs to generate CUDA kernels[3].

One of the most technically challenging but appealing aspects of research to me is being able to define and solve a new problem that changes how we view a field. I want to pursue a PhD at Columbia University because it uniquely offers the freedom, time, and resources to deeply understand and contribute to unsolved problems. The remainder of this statement serves as an outline of my research experiences so far, as well as future research directions I am interested in exploring.

Efficient ML: Scaling up and scaling down

I have generally been interested in the dual problem of enabling models to scale up efficiently, and finding the Pareto frontier of compute constraints and model performance. In addition to my crucial efficiency work on Ligo's open-source AlphaFold3[1], I have been working at Snapchat as a research intern on reducing the memory load of huge embedding tables in production recommendation systems, which can easily be billions of unique embeddings. After finishing my summer internship, I **proposed an idea about reducing the memory footprint of embedding tables** to Dr. Shah. I am now leading this project to investigate **empirical downwards-scaling laws for embedding table precision, dimension, and cardinality** on sequential recommendation tasks. Finally, I have been working on a long-term project with Professor Kai Li's ML systems group for over a year to accelerate deep learning and computational linear algebra workloads on heterogeneous clusters by minimizing redundant network traffic at the NCCL/socket level.

Evaluating model capabilities systematically

During my undergrad, I felt unconvinced that algorithms without provable guarantees were reliable enough to be used in production settings, so I became especially interested in how to

[1] https://github.com/Ligo-Biosciences/AlphaFold3?#msa-pair-averaging-efficiency
[2] https://alexzhang13.github.io/blog/2024/efficient-dl/
[3] https://gist.github.com/msaroufim/087c2a358c505e287a926e6a27b3e3b0

rigorously evaluate existing model capabilities. My first research project on video captioning models lacked any strong evaluation benchmarks, making it difficult to provide convincing evidence that our approach was robust. This work motivated [4], where I co-led the creation of the **first visual understanding benchmark with 1000+ hours of annotated footage to specifically evaluate long-horizon reasoning**. In addition to building the dataset, I found from our experiments that even on simple counting tasks, existing approaches struggled to localize relevant semantic information across long video streams. We published this work **in the ICLR 2024 Workshop on Data-centric Machine Learning Research**. Very recently, in [2], I co-authored a benchmark to **evaluate multimodal coding agents on real software engineering pull requests**, which is arguably the most realistic setting for automated coding agents. In addition to building the dataset, I led the experiments on this work and found that pre-existing agent scaffolds on prior software-engineering agentic tasks are over-engineered for Python repositories and unable to leverage non-text modalities. This work is **under submission at ICLR 2025**.

Language-grounded agents.

Fundamentally, I believe that **language is the most human-intuitive interface** for communicating our intent with artificial agents. I first approached Professor Narasimhan with an ambitious idea about **leveraging strategy guides to train reinforcement learning agents**, which are notoriously sample inefficient. This idea eventually transformed into [1], where we disentangled an agent's understanding of the world from its policy and showed that this "world model" could be grounded in language. As the lead on this project, I wanted to focus on evaluating "**compositional generalization**", so I also built a grid-world task to probe the agent's ability to decompose language semantics and the observations that ground them. Surprisingly, our agent can solve tasks in **environments with completely unseen entities using only language descriptions**, and we published the paper at the **ACL 2024 Workshop on Spatial Language Understanding and Grounded Communication for Robotics**.

The Future.

My past experiences and current research interests lie in the intersection of hardware-aware algorithms, rigorous theory, and generative models for natural language applications. For this reason, I am particularly excited about working on researching efficient methods for more scalable and accessible AI, stronger evaluations for understanding our models, and applications of language models to tasks like coding. I am also interested in applications of AI to the sciences, as well as theoretical research for understanding model dynamics and convergence behavior. At Columbia, I want to develop a wide foundation of skills from the math / natural sciences down to the hardware that I can apply to my research – I strongly believe that many ideas come from intuition in other fields, and I hope my research centers on this belief on my path to becoming a professor.

References

[1] **Alex Zhang***, Khanh Nguyen*, Jens Tuyls, Albert Lin, and Karthik Narasimhan. Language-guided World Models: A Model-based Approach to AI Control. In *Proceedings of the 4th Workshop on Spatial Language Understanding and Grounded Communication for Robotics (SpLU-RoboNLP at ACL 2024)*, pp. 1–16, 2024.

[2] John Yang*, Carlos E. Jimenez*, **Alex Zhang**, Kilian Lieret, Joyce Yang, Xindi Wu, Ori Press, Niklas Muennighoff, Gabriel Synnaeve, Karthik R. Narasimhan, Diyi Yang, Sida I. Wang,

and Ofir Press. SWE-bench Multimodal: Do AI Systems Generalize to Visual Software Domains? *arXiv preprint, submitted to ICLR 2025*, 2024.

[3] **Alex Zhang** and Karthik Narasimhan. Leveraging Language to Enhance Decision-making AI. *Princeton University Data Space (Outstanding Thesis Award)*, 2024.

[4] Aniket Agarwal*, **Alex Zhang***, Karthik Narasimhan, Igor Gilitschenski, Vishvak Murahari, and Yash Kant. Building Scalable Video Understanding Benchmarks through Sports. *DMLR Workshop (ICLR 2024)*, 2023.

[5] Michael Tang* and **Alex Zhang***. Transaction Fee Mining and Mechanism Design. *arXiv preprint*, 2023.